Finding Influential Cases in Linear Regression: A Review

by

R. Dennis Cook and Sanford Weisberg

Technical Report No. 338
February 1, 1979

University of Minnesota
School of Statistics
St. Paul, Minnesota 55108

*Work supported by grant 1-R01-GM25587-01 from the National Institute of General Medical Science, NIH. We are grateful to G.W. Stewart for several enlightening discussions concerning computational problems relevant to this paper.

Table of Contents

Abstract

Traditionally, most of the effort in fitting full rank linear regression models has centered on the study of the presence, strength and form of relationships between the measured variables. As is now well-known, least squares regression computations can be strongly influenced by a few of the cases in a data set, and a fitted model may more accurately reflect unusual features of those cases rather than the overall relationships between the variables. It is of interest, therefore, for an analyst to be able to find influential cases, and, based on them, make decisions concerning their usefulness in a problem at hand.

Based on an empirical influence function, we review methodologies for studying the influence of individual cases on a regression problem and put a number of possible measures that have been proposed into a general framework. We then consider the study of the simultaneous influence of several cases, including important computational considerations. Several additional issues are also discussed. We conclude with a carefully worked example, using data from the Florida Area Cumulus Experiments (FACE) on cloud seeding.

Key words: Linear models; influence functions; distance measures; robustness; ridge regression; residual plotting; outlier tests; cloud seeding.

1. Introduction

The problems we consider arise in the context of the linear model

$$Y = X\beta + e \tag{1.1}$$

where Y is an nxl vector of responses, X is an nxp matrix of known constants, β is a pxl parameter vector, and e is an nxl vector of errors. Data analyses based on this model usually center on the presence, form and strength of relationship between the response and independent variables (columns of X). Estimation, hypothesis testing, model selection and prediction are typical concerns.

Recently, interest in the role that individual observations or cases play in controlling an analysis has increased. (Here, observation or case refers to an individual response, y_i, in combination with the associated design point, or row of X.) Individual cases or goups of cases can exert a substantial influence on the analysis and yet go undetected when the residuals are examined. A case may be judged "influential" if important features of the analysis are altered substantially when the case is deleted. Clearly, failure to detect such cases can result in severe loss of information. For example, in a clinical trial, one patient with background that is different from that of the bulk of the patients may have a large effect on estimation and hence results may reflect this one case, and not accurately portray the data as a whole. Also, investigations of the causes of influential cases may lead to fresh insights into experimental methodologies. For example, an influential observation may be an indication of a region in the independent variable space with inadequate coverage.

An important tool in the study of influential cases is an empirical version of Hampel's influence function (1974). The influence function itself may be used to portray the potential effect of possible design points on estimators. Robust methods have been proposed that lead to bounded influence functions (e.g. Andrews, et al., 1972). Our use of the empirical influence function is somewhat different. We consider the least squares estimator $\hat{\beta} = (X^TX)^{-1}X^TY$ as a fixed point and then, given the observed data, we ask how an alternative estimator of β would compare to $\hat{\beta}$ if either the data were modified, or the estimation method were changed. In this way the influence function is used to study the sensitivity of the least squares estimator rather than to find alternative estimation techniques.

Let $\hat{\beta}_A$ be an estimator of β based either on a modified data set or an alternative estimation technique (e.g., ridge regression). The <u>empirical influence function</u>, I_A, is defined to be

$$I_A = \hat{\beta}_A - \hat{\beta} \tag{1.2}$$

In general, I_A is a p-vector, and functions of I_A of lower dimension can be profitably studied. For the moment, assume that interest centers on $\hat{\beta}$ itself, rather than some function of it (such as a single component, a prediction at a point, etc.). For some inner product matrix M and scale factor c, the <u>distance</u>, $D_A(M,c)$, between $\hat{\beta}_A$ and $\hat{\beta}$ is taken to be

$$D_A(M,c) = \frac{I_A^T M I_A}{c} \tag{1.3}$$

In this paper, we study (1.3), and statistics derived from it, for varying choices of A, M, and c. The most fruitful approach we have found is in looking

at the effects of deleting single cases from the data one at a time, as reviewed in Section 2. In Section 3, we proceed to the problem of studying the effects of simultaneous deletion of several cases. Section 4 contains methods for studying regression problems when a linear transformation of β is of interest. Section 5 contains a discussion of ridge regression, and Section 6 contains some comments on computation. Finally, in the last section we present an example in some detail, using data from the Florida Area Cumulus Experiments (Woodley, et al. 1977) on cloud seeding.

2. Deleting One Case at a Time

In this section, we review statistics for studying the impact on $\hat{\beta}$ of individual cases taken one at a time. We assume the model (1.1) with rank(X) = p, Cov(e) = $\sigma^2 I$. The least squares estimator of β using the full data is therefore $\hat{\beta} = (X^TX)^{-1}X^TY$; the full sample estimate of σ^2 is $s^2 = Y^T(I-X(X^TX)^{-1}X^T)Y/(n-p)$. One measure of the impact of the i-th case is simply the i-th residual $r_i = y_i - x_i^T\hat{\beta}$, where x_i^T is the i-th row of X. Very loosely speaking, cases with large r_i have been considered as ones for which the model fails either due to incorrect functional form or because of an outlier in y. As we shall see, this notion can be formalized through the use of empirical influence functions.

We shall need some additional notations: A subscript "$_{(i)}$" added to a quantity means "with the i-th case deleted." Thus, for example, $X_{(i)}$ is an (n-1) x p matrix derived from X by deleting the i-th row X_i^T, $\hat{\beta}_{(i)} = (X_{(i)}^TX_{(i)})^{-1}X_{(i)}^TY_{(i)}$, etc. Also of importance is the projection matrix $V = X(X^TX)^{-1}X^T$, an n x n rank p matrix that projects onto the column space of X. Elements of V are denoted v_{ij}; in this section, the diagonal entries v_{ii} are of special interest.

The empirical influence function (1.2) for $\hat{\beta}_{(i)}$ is given by

$$I_i = (\hat{\beta}_{(i)} - \hat{\beta}) \tag{2.1}$$

and the distance function (1.3) is given, for some M and c, by

$$D_i(M,c) = \frac{(\hat{\beta}_{(i)}-\hat{\beta})^TM(\hat{\beta}_{(i)}-\hat{\beta})}{c} \tag{2.2}$$

"Large" values of $D_i(M,c)$ would correspond to cases that, when deleted, result in large movement in the estimate of β. The problem of defining "large" will be addressed shortly.

We shall call a case with a large value of $D_i(M,c)$ <u>influential for estimating</u> β relative to (M,c).

One natural choice for M and c are (X^TX) and ps^2 respectively, although others are possible. The resulting statistic, suggested by Cook (1977) is

$$D_i((X^TX), ps^2) = \frac{(\hat{\beta}_{(i)}-\hat{\beta})(X^TX)(\hat{\beta}_{(i)}-\hat{\beta})}{ps^2} \qquad (2.3)$$

Form (2.3) has a useful geometric interpretation, as the magnitude of the distance between $\hat{\beta}$ and $\hat{\beta}_{(i)}$ may be assessed by comparing $D_i(X^TX,ps^2)$ to the probability points of the central F with p and n-p degrees of freedom. This is equivalent to studying the least squares confidence ellipsoids for β based on the full data, and finding the ellipsoid that passes through $\hat{\beta}_{(i)}$. The F distribution is used only to transform $D_i(X^TX, ps^2)$ to a more familiar scale.

Also, (2.3) may be rewritten in the form

$$D_i((X^TX), ps^2) = \frac{(\hat{Y}_{(i)}- \hat{Y})^T(\hat{Y}_{(i)}- \hat{Y})}{ps^2}$$

(Bingham 1977) suggesting that, in the estimation space, $D_i((X^TX), ps^2)$ is, aside from the scale factor ps^2, the ordinary squared Euclidean distance that the fitted vector moves when the i-th case is deleted from the data.

Because of the relationship between $\hat{\beta}$ and $\hat{\beta}_{(i)}$, computing and using the $D_i(X^TX, ps^2)$ can be remarkably easy. All of the necessary results can be derived as a special case of the Sherman-Morrison-Woodbury formula:

$$(X^T_{(i)}X_{(i)})^{-1} = (X^TX)^{-1} + \frac{(X^TX)^{-1}x_i x_i^T(X^TX)^{-1}}{1-v_{ii}} \qquad (2.4)$$

which can be used to show (Cook (1977); Bingham (1977))

$$\hat{\beta}-\hat{\beta}_{(i)} = \frac{(X^TX)^{-1}x_i r_i}{1-v_{ii}}$$

It follows immediately that

$$D_i(X^TX,\ ps^2) = \frac{1}{p}\left(\frac{r_i}{s\sqrt{1-v_{ii}}}\right)^2\left(\frac{v_{ii}}{1-v_{ii}}\right) . \tag{2.5}$$

We define $t_i = r_i/s\sqrt{1-v_{ii}}$ to be the studentized residual since $var(r_i) = \sigma^2(1-v_{ii})$. Thus,

$$D_i(X^TX,\ ps^2) = \frac{1}{p}\left(t_i^2\right)\left(\frac{v_{ii}}{1-v_{ii}}\right) \tag{2.6}$$

$D_i(X^TX,\ ps^2)$ is a product of three terms from the full data: a scale factor p^{-1}, a residual term t_i^2, and $v_{ii}/(1-v_{ii})$. As each of these last two is important, we discuss them separately.

2.1 Looking at the v_{ii}. The term $v_{ii}/(1-v_{ii})$ in (2.6) was interpreted by Cook (1977) as the ratio $Var(\hat{y}_i)/Var(r_i)$, a measure of the relative precision of estimation at x_i. Hoaglin and Welsch (1978) have also discussed the v_{ii} (which they call h_i) in some detail. Cases with large values of v_{ii} are potentially influential cases since $v_{ii}/(1-v_{ii})$ will be large. Hoaglin and Welsch call cases with large values of v_{ii} leverage points, and we follow their example:

A case with a large value of v_{ii} is called a high leverage case.

The v_{ii} can be usefully viewed as being a distance measure in the X-space, since v_{ii} measures the distance from x_i to $\bar{x}$ (or the origin if regression is through the origin) relative to the inner product X^TX. Using (2.4), one can show that

$$x_i^T(X_{(i)}^TX_{(i)})^{-1}x_i = \frac{v_{ii}}{1-v_{ii}} \tag{2.7}$$

Thus, $v_{ii}/(1-v_{ii})$ is the distance from x_i to the center of the remaining n-1 cases in the sample. The v_{ii} are also related to the Mahalanobis distance, MD_i, by $MD_i = (n-1)(v_{ii}-1/n)$. Adding the assumption of multivariate normality, Welsch and Kuh (1977) note that $(n-p)(v_{ii}-1/n)/(p-1)(1-v_{ii})$ is distributed as F with (p-1, n-p) degrees of freedom.

The term $v_{ii}/(1-v_{ii})$ may be interpreted also as the total change in the variance of prediction at the points x_j, $j = 1,\ldots n$, when x_i is deleted,

$$v_{ii}/(1-v_{ii}) = [\sum_j \mathrm{Var}(x_j^T\hat{\beta}) - \sum_j \mathrm{Var}(x_j^T\hat{\beta}_{(i)})]/\sigma^2 \quad .$$

This may be verified by first noting that

$$\frac{1}{\sigma^2}\mathrm{Var}(x_j^T\hat{\beta} - x_j^T\hat{\beta}_{(i)}) = v_{ij}^2/(1-v_{ii}) = [\mathrm{Var}(x_j^T\hat{\beta}) - \mathrm{Var}(x_j^T\hat{\beta}_{(i)})]/\sigma^2$$

Next, summing over $j = 1,\ldots n$ and using the property that $\sum_j v_{ij}^2 = v_{ii}$ produces the desired result.

Other properties of the v_{ii} can be derived from the fact that V is a projection matrix. Since rank(V) = p = trace(V) = Σv_{ii}=p. Also $1/n \leq v_{ii} \leq 1/c$ where c is the number of rows in X that are identical to x_i^T. Assuming that the intercept is in the model, and $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_{p-1}$ are the eigenvalues of the corrected cross product matrix for the data, and $P_1, \ldots, P_{p-1}$ are the corresponding eigenvectors, then by the spectral decomposition of the corrected cross product matrix,

$$v_{ii} = \frac{1}{n} + \sum_{\ell=1}^{p-1} \left(\frac{P_\ell^T(x_i - \bar{x})}{\sqrt{\mu_\ell}} \right)^2 \tag{2.8}$$

Writing $\theta_{P_\ell x_i}$ as the angle between P_ℓ and $(x_i - \bar{x})$, then

$$\cos(\theta_{P_\ell x_i}) = \frac{P_\ell^T(x_i - \bar{x})}{(x_i - \bar{x})^T(x_i - \bar{x})} \tag{2.9}$$

and

$$v_{ii} = \frac{1}{n} + (x_i - \bar{x})^T(x_i - \bar{x}) \sum_{\ell=1}^{k-1} \frac{\cos^2\theta_{P_\ell x_i}}{\mu_\ell} \tag{2.10}$$

Thus, v_{ii} is large if (1) x_i is far from $\bar{x}$ that is; it is well removed from the bulk of the cases, and (2) x_i is substantially in a direction of an eigenvector corresponding to a very small eigenvalue of the corrected cross product matrix. On the other hand, if $(x_i - \bar{x})$ is small, v_{ii} will be small regardless of its direction. Contours of constant v_{ii} are ellipsoids centered at $\bar{x}$ with axes given by the eigenstructure of X^TX.

2.2 Studentized residuals. The studentized residuals are a marked improvement over the r_i because they are insensitive to variations in $Var(e_i) = \sigma^2(1-v_{ii})$ if the model fit is correct (Behnken and Draper 1972). Also, the t_i are a monotonic transformation of the likelihood ratio test that the i-th case is an outlier (Cook 1979). Specifically, in the model

$$Y = X\beta + \theta u_i + e \tag{2.11}$$

where u_i is the i-th unit vector, the normal theory likelihood ratio test for $\theta = 0$ is given by

$$F_i = \frac{r_i^2}{s_{(i)}^2(1-v_{ii})} \quad t_i^2\left(\frac{n-p-1}{n-p-t_i^2}\right) \tag{2.12}$$

where, under normality, $F_i \sim F(1, n-p-1)$, ignoring the multiple test problem. However, the non-centrality parameter of F_i given $\theta \neq 0$ is given by $\eta = \theta^2(1-v_{ii})/2\sigma^2$ (Cook 1979), and outliers with large v_{ii}, e.g. at unusual points in the sample space, will be relatively difficult to detect. Since v_{ii} is an increasing function of p, as the size of a model increases, the power of F_i will decrease.

A case will be called a potential outlier if F_i (or equivalently, t_i) is large.

2.3. Residual Plots. Because the variances of the t_i are all the same (if the model is correct, with no outliers), residual plots using the t_i rather than the r_i are more readily interpretable, since the plotted values are all on the same scale. The only disadvantage to using the studentized residuals is that they are slightly correlated with the fitted values $\hat{y}_i$. However, in our experience this slight correlation is negligible.

2.4 Discussion. The distance measure (2.6) measures the total impact of the i-th case on the estimation. A point near the center of the data (v_{ii} small) is unlikely to be influential even if the case is a potential outlier. On the other hand, leverage cases are likely to be influential whether or not the case is a potential outlier, since $v_{ii}/(1-v_{ii})$ is unbounded, and t_i^2 is a random variable with mean near 1 if the i-th case is not an outlier and mean generally greater than 1 if it is an outlier. Clearly, a complete analysis requires computation of three quantities for each case: t_i^2 (or F_i); v_{ii} (or $v_{ii}/(1-v_{ii})$) and $D_i(X^TX, ps^2)$.

Since only outliers corresponding to cases with large values of v_{ii} will have significant impact on the estimation of β, in testing for outliers we may wish to take advantage of our relative lack of interest in cases with small values of v_{ii}. For example, if the Bonferroni inequality is applied to the F distribution to get significance levels for the outlier test, then we may choose to apportion probability of errors unequally, giving smaller critical values to cases with larger values of v_{ii}, and larger critical values to cases with small v_{ii}. This is mathematically permissible since the distribution of the outlier test depends on the random distribution of the vector Y, not on the (assumed fixed) values of the v_{ii}. In large samples, this method, when formalized, may result in substantial increases in power for the outlier test at the high leverage points.

2.5. Alternative choices of M, c. In the general form for the distance measure (2.2), the choice of M and c was left arbitrary, as obviously many options other than $M = X^TX$, $c = ps^2$ are reasonable. Table 1 lists some choices, and the reference, if any, where the alternative was suggested. Although no detailed comparison for the differing M and c has been made, it is likely that any choice for M, c such that $D_i(M,c)$ is location/scale invariant (excluding, therefore $M=I$, $c=ps^2$) would give approximately the same information.

The last line of Table 1 gives a statistic proposed by Andrews and Pregebon (1978). They have considered the problem of looking at the importance of the i-th case in a different way. The deter-

minant of a cross product matrix is an important criterion in experimental designs, with large values being associated with good designs, since the square root of the determinant is proportional to the inverse of the volume of the resulting confidence ellipsoid. With this in mind, they adjoin Y to X and define $X^* = (Y\ X)$ and $X^*_{(i)} = (Y_{(i)}\ X_{(i)})$. Then they suggest looking at the ratio

$$R_i(X^*) = \frac{\det(X^{*T}_{(i)}\ X^*_{(i)})}{\det(X^{*T}\ X^*)} \quad . \tag{2.13}$$

$R_i(X^*)$ corresponds to the proportion of the total volume generated by X^* that is not due to the i-th case. "Distant" or unusual sets of cases will tend to have $R_i(X^*)$ small, while cases in the "middle" of the data will account for little volume. $R_i(X^*)$ will measure information similar to the distance measures based on (1.2). It should be noted, however, that (2.13) is invariant with respect to specification of the response variable Y.

In the sequel, only the distance measure $D_i(X^TX,\ ps^2)$ will be discussed in detail, since the other reasonable choices for $(M,\ c)$ lead to essentially analogous statistics. For simplicity of notation, we shall write D_i for $D_i(X^TX,\ ps^2)$.

Table 1.

$$D_i(M, C) = (\hat{\beta}_{(i)} - \hat{\beta})^T M(\hat{\beta}_{(i)} - \hat{\beta})/c$$

Measure	M	c	Reduced form	Reference
$(\text{NDFBETAS})^2$	$[\text{diag}(X^TX)^{-1}]^{-1}$	$s^2_{(i)}$	$\frac{n-p}{p} F_i \frac{x_i^T(X^TX)^{-1}M(X^TX)^{-1}x_i}{1-v_{ii}}$	Welsch and Kuh (1977)
$(\text{DFFITS})^2$	X^TX	$s^2_{(i)}$	$\frac{n-p}{p} F_i \frac{v_{ii}}{1-v_{ii}}$	Welsch and Kuh (1977)
D_i	X^TX	s^2	$\frac{1}{p} t_i^2 \frac{v_{ii}}{1-v_{ii}}$	Cook (1977)
	$X^T_{(i)}X_{(i)}$	s^2	$\frac{1}{p} t_i^2 v_{ii}$	
	$X^T_{(i)}X_{(i)}$	$s^2_{(i)}$	$\frac{1}{p} F_i v_{ii}$	
	I	s^2	$t_i^2 \frac{x_i^T(X^TX)^{-2}x_i}{1-v_{ii}}$	
	—	-	$\left(\frac{n-p-t_i^2}{n-p}\right)(1-v_{ii})$	Andrews and Pregibon (1978)

3. Many cases at a time.

The one at a time statistics can be expected to provide the majority of the information needed to carry out the analysis needed. However, in some data sets, subsets of cases can be jointly influential, but if taken one at a time the cases are not influential. Consider, for example, Figures 1a and 1b. In Fig. 1a, if point A or point B were deleted, the fitted model may not change very much. If both are deleted, however, estimates of parameters may show severe changes Conversely, in Fig. 1b, if C or D is deleted, the fitted line will change; if C and D are both deleted, the line will stay about the same.

The generalization of the distance measure for deleting several cases requires some additional notation and some algebra. Let the m-vector $\underset{\sim}{i}$ index a set of m cases that are to be deleted. The subscript "$_{(\underset{\sim}{i})}$" will mean "with the m cases indexed in $\underset{\sim}{i}$ deleted," and "$_{\underset{\sim}{i}}$" without parentheses will mean with only the cases indexed by $\underset{\sim}{i}$ remaining. So, for example, $V_{\underset{\sim}{i}}$ will be the $m \times m$ submatrix of V formed by the intersection of the rows and columns indexed in $\underset{\sim}{i}$, and $r_{\underset{\sim}{i}}$ is the $m \times 1$ vector of residuals for the cases indexed in $\underset{\sim}{i}$. The empirical influence function is

$$I_{\underset{\sim}{i}} = \hat{\beta}_{(\underset{\sim}{i})} - \hat{\beta} \tag{3.1}$$

and the distance function $D_{\underset{\sim}{i}}(X^TX, ps^2) = D_{\underset{\sim}{i}}$ is

$$D_{\underset{\sim}{i}} = \frac{(\hat{\beta}_{(\underset{\sim}{i})} - \hat{\beta})^T (X^TX) (\hat{\beta}_{(\underset{\sim}{i})} - \hat{\beta})}{ps^2} \tag{3.2}$$

The geometric interpretation of $D_{\underset{\sim}{i}}$ is identical to that of D_i. An influential subset for estimating β would have $D_{\underset{\sim}{i}}$ large.

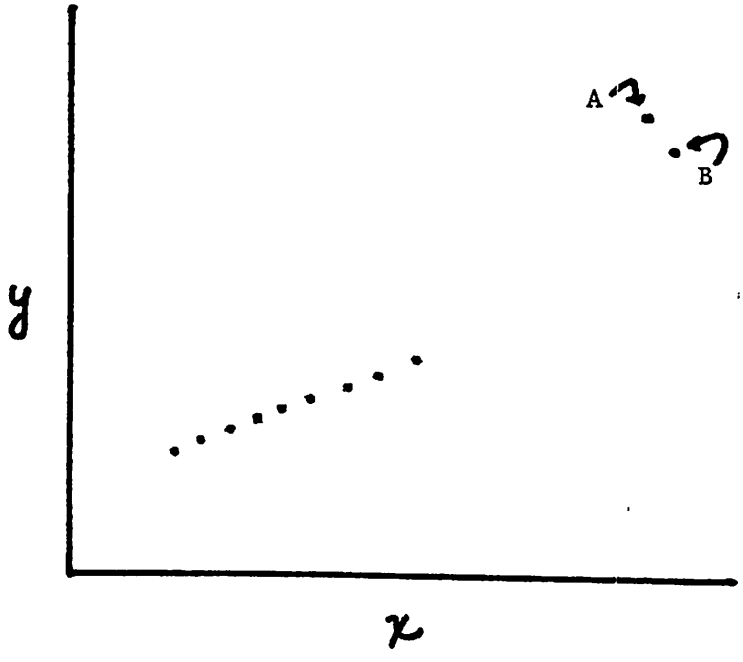


Figure 1a. Cases A and B are jointly influential, but not individually influential.

Figure 1b. Cases C and D are singly influential but not jointly influential.

A convenient formula for $D_{\underset{\sim}{i}}$ can be derived from the result (Bingham (1977))

$$\hat{\beta}_{(\underset{\sim}{i})} - \hat{\beta} = -(X^TX)^{-1}X_{\underset{\sim}{i}}^T(I-V_{\underset{\sim}{i}})^{-1}r_{\underset{\sim}{i}}$$

so that

$$D_{\underset{\sim}{i}} = \frac{r_{\underset{\sim}{i}}^T(I-V_{\underset{\sim}{i}})^{-1}V_{\underset{\sim}{i}}(I-V_{\underset{\sim}{i}})^{-1}r_{\underset{\sim}{i}}}{ps^2} \tag{3.4}$$

Methods for computing (3.4) are discussed in Section 6.

Further insight into $D_{\underset{\sim}{i}}$ is obtained by applying the spectral decomposition to $V_{\underset{\sim}{i}}$: There is an $m \times m$ orthoginal matrix Γ and an $m \times m$ diagonal matrix $\Lambda = \text{diag}\,(\lambda_1, \ldots, \lambda_m)$, with $0 \leq \lambda_1 \leq \ldots \leq \lambda_m \leq 1$, such that

$$V_{\underset{\sim}{i}} = \Gamma^T\Lambda\Gamma \tag{3.5}$$

The two cases $\lambda_m = 1$ and $\lambda_m < 1$ must be treated separately. If $\lambda_m = 1$ then the inverses in (3.4) do not exist, and, if the cases indexed by $\underset{\sim}{i}$ are removed from the data, the resulting data is rank deficient, and a unique estimator $\hat{\beta}_{(\underset{\sim}{i})}$ does not exist. In a sense, therefore, if $\lambda_m = 1$, $D_{\underset{\sim}{i}} = \infty$. If $\lambda_m < 1$, then we can rewrite (3.4) as

$$D_{\underset{\sim}{i}} = \frac{r_{\underset{\sim}{i}}^T(\Gamma^T\Gamma - \Gamma^T\Lambda\Gamma)^{-1}\Gamma^T\Lambda\Gamma(\Gamma^T\Gamma - \Gamma^T\Lambda\Gamma)^{-1}r_{\underset{\sim}{i}}}{ps^2} \tag{3.6}$$

$$= \frac{(\Gamma r_{\underset{\sim}{i}})^T(I-\Lambda)^{-1}\Lambda(I-\Lambda)^{-1}(\Gamma r_{\underset{\sim}{i}})}{ps^2}.$$

Letting $g^T = (g_1, \ldots, g_m) = \Gamma r_{\underset{\sim}{i}}$

$$D_{\underset{\sim}{i}} = \frac{g^T(I-\Lambda)^{-1}\Lambda(I-\Lambda)^{-1}g}{ps^2} \tag{3.7}$$

$$= \frac{\sum_{\ell=1}^{m} g_\ell^2 \frac{\lambda_\ell}{(1-\lambda_\ell)^2}}{ps^2}$$

Now, each g_ℓ is a linear combination of the elements of $r_{\underset{\sim}{i}}$, and $\mathrm{Var}(g) = \mathrm{Var}(\Gamma r_i) = \sigma^2\Gamma\Gamma^T(I-\Lambda)\Gamma\Gamma^T = \sigma^2(I-\Lambda)$. Thus, the g_ℓ are uncorrelated and $\mathrm{var}(g_\ell) = \sigma^2(1-\lambda_\ell)$. Setting

$$h_\ell^2 = \frac{g_\ell^2}{s^2(1-\lambda_\ell)} \tag{3.8}$$

each h_ℓ is identically distributed . Then (3.7) may be rewritten

$$D_{\underset{\sim}{i}} = \frac{1}{p}\sum_{\ell=1}^{m} h_\ell^2 \frac{\lambda_\ell}{1-\lambda_\ell} \tag{3.9}$$

The resemblance of (3.9) to (2.6) is striking. The role of the t_i^2 is assumed by the h_ℓ^2 and $v_{ii}/(1-v_{ii})$ are replaced by the $\lambda_\ell/(1-\lambda_\ell)$. In (3.9), a sum over m orthogonal directions is required; in (2.6), $m = 1$.

One generalization of the squared studentized residuals to m cases is given by $r_{\underset{\sim}{i}}^T (I-V_{\underset{\sim}{i}})^{-1} r_{\underset{\sim}{i}}/s^2 = \Sigma h_\ell^2$. A likelihood ratio test that the m cases are an outlying set, under normal theory (Gentleman and Wilk (1975)) can be computed as

$$F_{\underset{\sim}{i}} = \frac{(n-p-m)\ \Sigma h_\ell^2}{n-p-m\ \Sigma h_\ell^2} \tag{3.10}$$

The nominal distribution of $F_{\underset{\sim}{i}}$ is $F(m, n-p-m)$.

The information contained in the $\lambda_\ell/(1-\lambda_\ell)$ is measured by the heterogeniety in the eigenvalues $\lambda_1, \ldots, \lambda_m$. A simple summary for this heterogeniety is the condition number of $V_{\underset{\sim}{i}}$, defined by $\kappa_{\underset{\sim}{i}} = \sqrt{\lambda_m/\lambda_1}$. Although this quantity

is somewhat less informative than all of the $\lambda_\ell/(1-\lambda_\ell)$, the condition number may be easier to compute and interpret. Subsets with $\kappa_{\underset{\sim}{i}}$ large are potential leverage subsets.

3.1 Looking at the $D_{\underset{\sim}{i}}$.

The principal goal in examining subsets of $m > 1$ cases is to find groups of cases that, while not influential individually, are influential when taken as a group. Finding influential subsets that include cases that are individually influential adds little information because the observed influence of the subset will be due in part to the influence of the single influential case. Conversely, finding an uninfluential subset that includes one or more cases that are singly influential would not decrease the interest in those cases. Thus, candidates for inclusion in subsets will have small values of D_i , but they may well have relatively large values of v_{ii} or t_i .

Because of the enormous amount of computing necessary for finding the subsets with large $D_{\underset{\sim}{i}}$, a systemmatic approach is required. We assume that the one at a time statistics v_{ii} and r_i are available in computer memory while the off diagonal elements of V, the V_{ij} are not. We can then compute an upper bound for $D_{\underset{\sim}{i}}$, and only if this bound is sufficiently large must the $D_{\underset{\sim}{i}}$ be computed exactly.

For the first upper bound, since $\lambda_m/(1-\lambda_m)^2 \geq \lambda_\ell/(1-\lambda_\ell)^2 \quad \ell = 1,\ldots,m$, (3.7) can be approximated by

$$D_{\underset{\sim}{i}} \leq \frac{1}{ps^2} \frac{\lambda_m}{(1-\lambda_m)^2} \sum_{\ell=1}^{m} g_\ell^2 \ .$$

But $\Sigma g_\ell^2 = (r_{\underset{\sim}{i}}^T \Gamma^T \Gamma r_{\underset{\sim}{i}}) = r_{\underset{\sim}{i}}^T r_{\underset{\sim}{i}} = \sum_{i \varepsilon \underset{\sim}{i}} r_i^2$. Hence

$$D_{\underset{\sim}{i}} \leq \frac{\lambda_m}{(1 - \lambda_m)^2} \left(\frac{\sum_{i \in \underset{\sim}{i}} r_i^2}{ps^2} \right) . \tag{3.11}$$

For (3.11) to be useful, λ_m must be replaced by an approximation that can be computed without need for explicitly finding λ_m. The easiest approximation to use is $\lambda_m \leq tr(V_{\underset{\sim}{i}})$ assuming $tr(V_{\underset{\sim}{i}}) \leq 1$, so that

$$D_{\underset{\sim}{i}} \leq \frac{tr(V_{\underset{\sim}{i}})}{(1 - tr(V_{\underset{\sim}{i}}))^2} \left(\frac{\sum_{i \in \underset{\sim}{i}} r_i^2}{ps^2} \right)$$

or, equivalently

$$D_{\underset{\sim}{i}} \leq \frac{\sum_{i \in \underset{\sim}{i}} v_{ii}}{(1 - \sum_{i \in \underset{\sim}{i}} v_{ii})^2} \frac{\sum_{i \in \underset{\sim}{i}} r_i^2}{ps^2} . \tag{3.12}$$

Approximation (3.12) depends only on the one at a time statistics, giving a potentially different upper bound for each $\underset{\sim}{i}$. For any subset with $tr(V_{\underset{\sim}{i}}) \geq 1$, a better approximation to λ_m is required. This, in turn, requires formation of $V_{\underset{\sim}{i}}$. If m is small (2 or 3) exact computation of $D_{\underset{\sim}{i}}$ is probably as efficient as obtaining an approximation for λ_m.

For a fixed m, let $T = \max_{\underset{\sim}{i}} (\sum_{i \in \underset{\sim}{i}} v_{ii})$ and $R^2 = \max_{\underset{\sim}{i}} \sum_{i \in \underset{\sim}{i}} r_i^2$, wherein each, $\underset{\sim}{i}$ varies over all subsets of m under condiseration. Two upper bounds for the right side of (3.12) are then

$$D_{\underset{\sim}{i}} \leq \frac{tr(V_{\underset{\sim}{i}})}{(1 - tr(V_{\underset{\sim}{i}}))^2} \frac{R^2}{ps^2} \tag{3.13}$$

and, if $T < 1$,

$$D_{\underset{\sim}{i}} \leq \frac{T}{(1 - T)^2} \frac{\sum r_i^2}{ps^2} . \tag{3.14}$$

These last two may be combined to give

$$D_{\underset{\sim}{i}} \leq \frac{T}{(1-T)^2} \frac{R^2}{ps^2} \quad . \tag{3.15}$$

Note that (3.12) $\leq$ (3.13) $\leq$ (3.15) , and (3.12) $\leq$ (3.14) $\leq$ (3.15) . If $m = 1$, all four approximations are exact.

These four approximations can be used to study systematically all subsets of size m , by first choosing a cutoff D^* , perhaps $D^* = F(.50; p,n-p)$. The four inequalities can be applied in the order (3.15), then (3.14) or (3.13), and then (3.12). If (3.12) is not satisfied, then $D_{\underset{\sim}{i}}$ must be computed exactly. These computations will be facilitated by ordering the v_{ii} and the t_i , largest to smallest for each.

4. Linear Combinations

In this section we extend the previous discussion to accomodate the situation in which q linearly independent combinations of the elements of β are of interest. This may be desirable when, for example, interest centers on a selected subset of β. Also, once an influential observation has been found using D_i it may be desirable to isolate the effects on the individual components of $\hat{\beta}$.

Let $\hat{\psi} = L\hat{\beta}$, where L is a $q \times p$ rank q matrix. The distance, $D_{\underset{\sim}{i}}(\psi)$, between $\hat{\psi}$ and $\hat{\psi}_{(\underset{\sim}{i})} = L\hat{\beta}_{(\underset{\sim}{i})}$ is defined to be

$$D_{\underset{\sim}{i}}(\psi) = \frac{(\hat{\psi} - \hat{\psi}_{(\underset{\sim}{i})})^T [L(X^TX)^{-1}L^T]^{-1}(\hat{\psi} - \hat{\psi}_{(\underset{\sim}{i})})}{qs^2} \quad . \tag{4.1}$$

Note that this is a special case of the distance function $D_{\underset{\sim}{i}}(M, c)$ obtained by choosing $c = qs^2$ and

$$M = L^T[L(X^TX)^{-1}L^T]^{-1}L \quad . \tag{4.2}$$

Bingham (1977) has shown that the numerator of $D_i(\psi)$ can be written as

$$qs^2 D_{\underset{\sim}{i}}(\psi) = r_{\underset{\sim}{i}}^T(I - V_{\underset{\sim}{i}})^{-1}X_{\underset{\sim}{i}}(X^TX)^{-1}M(X^TX)^{-1}X_{\underset{\sim}{i}}^T(I - V_{\underset{\sim}{i}})^{-1}r_{\underset{\sim}{i}}^T \tag{4.3}$$

where M is given by (4.2). Apparently, further simplification is not possible without additional contraints on L. However, direct computation of $D_i(\psi)$ will probably be unnecessary for most $\underset{\sim}{i}$: Since $(X^TX)^{-\frac{1}{2}}M(X^TX)^{-\frac{1}{2}}$ is idempotent it follows from (4.3) that $qs^2 D_{\underset{\sim}{i}}(\psi) \leq ps^2 D_{\underset{\sim}{i}}$ and, therefore,

$$D_{\underset{\sim}{i}}(\psi) \leq \frac{p}{q} D_{\underset{\sim}{i}} \tag{4.4}$$

for all $\underset{\sim}{i}$ and ψ. Thus, if $D_{\underset{\sim}{i}}$ is negligible, $D_{\underset{\sim}{i}}(\psi)$ must be negligible also. The result (4.4) was mentioned by Cook (1979) for the case $m = 1$.

4.1 Predictions. If $q = 1$ and L is a vector of carrier values then $D_i(\psi)$ measures the distance between the prediction at L using the i-th case and the prediction at L without the i-th case and (4.3) simplifies to (Cook, 1977)

$$D_i(\psi) = pD_i\ \rho^2(x_i^T\hat{\beta},\ L\hat{\beta}) \tag{4.5}$$

where $\rho(\cdot, \cdot)$ denotes the correlation coefficient. When the i-th case is deleted the maximum movement (as measured by $D_i(\psi)$) in a prediction occurs at x_i. Of course, (4.5) applies also to situations in which a single component of β is of interest.

4.2. Subsets of β. Suppose without loss of generality that the last q components of β are of interest, and partition $X = (X_1\ X_2)$ where X_1 is $n \times (p-q)$ and X_2 is $n \times q$. Thus, $L = (0,\ I_q)$, and

$$(X^TX)^{-1}M(X^TX)^{-1} = (X^TX)^{-1} - \begin{pmatrix} (X_1^TX_1)^{-1} & 0 \\ \underset{\sim}{\ } & \\ 0 & 0 \end{pmatrix}$$

where M is given by (4.2). Substitution in (4.3) yields

$$qs^2D_{\underset{\sim}{i}}(\psi) = ps^2D_{\underset{\sim}{i}} - r_{\underset{\sim}{i}}^T(I - V_{\underset{\sim}{i}})U_{\underset{\sim}{i}}(I-V_{\underset{\sim}{i}})^{-1}r_{\underset{\sim}{i}}^T \tag{4.6}$$

where $U = X_1(X_1^TX_1)^{-1}X_1^T$, so $U_{\underset{\sim}{i}}$ is an $m \times m$ submatrix of U. Alternatively, (4.6) can be written as

$$qs^2D_i(\psi) = r_{\underset{\sim}{i}}^T(I - V_{\underset{\sim}{i}})^{-1}(V_{\underset{\sim}{i}} - U_{\underset{\sim}{i}})(I - V_{\underset{\sim}{i}})^{-1}r_{\underset{\sim}{i}}\ . \tag{4.7}$$

Thus, when a single case is deleted ($m = 1$)

$$D_i(\psi) = \frac{t_i^2}{qs^2} \frac{v_{ii} - u_{ii}}{1 - v_{ii}} \quad . \tag{4.8}$$

The influence of a single case on a selected subset of β may therefore be determined from the result of two separate regressions on the full data.

4.3 Ignoring the intercept. If interest centers on all coefficients except the intercept, then $p - q = 1$, $u_{ii} = 1/n$ and (4.8) reduces to

$$D_i(\psi) = \frac{t_i^2}{p - 1} \frac{v_{ii} - 1/n}{1 - v_{ii}} \quad . \tag{4.9}$$

Unless special importance is attached to the constant term or predictions are of interest, it may be desirable to use (4.9) as an exploratory tool for isolating influential observations, although there will be little difference between (4.9) and D_i for moderately large data sets.

Expression (4.9) can be written as

$$D_i(\psi) = \left[\frac{t_i^2}{n - p}\right]\left[\frac{(n - p)(v_{ii} - 1/n)}{(p - 1)(1 - v_{ii})}\right] \quad . \tag{4.10}$$

Recall that when the carriers are assumed to have a multivariate normal distribution the second factor on the right-hand-side of (4.10) is distributed as $F(p - 1, n - p)$ (Welsch and Kuh, 1977). In addition, it is easily verified that the first factor has a beta distribution with parameters $1/2$ and $(n - p - 1)/2$, $B(1/2, (n - p - 1)/2)$, and that the two factors are independent. It follows that the distribution of $D_i(\psi)$ when the intercept alone is ignored is the product of independent $B(1/2, (n - p - 1)/2)$ and $F(p - 1, n - p)$ random variables.

5. Ridge Regression

With some modification, the results of the previous sections may be applied directly to ridge estimation. In this section we suggest distance measures for comparing ridge and least squares estimations and for the detection of influential observations in ridge regresison. Only ordinary ridge regression is considered. The extension to generalized ridge regression is immediate, although somewhat more detailed.

The ordinary ridge estimate, $\tilde{\beta}$, of β for some $k \geq 0$ is given by

$$\tilde{\beta} = (X^TX + kI)^{-1}X^TY \quad . \tag{5.1}$$

Of course, if $k = 0$ then $\tilde{\beta} = \hat{\beta}$. Usually k is chosen as a function of the data; however, for the purposes of the discussion the manner in which k is chosen is irrelevant.

The ridge estimate may be considered a least squares estimate arising when the original data set is modified in a special way: Let $a_1, \ldots, a_p$ denote a set of p orthogonal vectors such that $a_i^Ta_i = k$. For example, without loss of generality take the a_i's to be $\sqrt{k}$ times the eigenvectors of X^TX. Let A denote a $p \times p$ matrix whose i-th row is a_i^T. It is straightforward to verify that $\tilde{\beta}$ is the least squares estimate of β from the fictitious model

$$Z = W\beta + e \tag{5.2}$$

where $Z^T = (Y^T, 0^T)$ and $W^T = (X^T, A^T)$. It follows immediately that all previous results hold with Y and X replaced by Z and W, respectively. However, in the case of ridge regression it is not clear that $M = W^TW$ is an appropriate inner product matrix.

5.1 Comparing Ridge and Least Squares Estimators. It seems natural to compare $\tilde{\beta}$ and $\hat{\beta}$ using $M = X^TX$ rather than $M = W^TW$ since the choice $M = X^TX$, $c = ps^2$ bases the comparison on the confidence ellipsoids associated with $\hat{\beta}$ while the appropriate interpretation for $M = W^TW$ is not apparent. Thus, we suggest using

$$D_{R,k} = \frac{(\tilde{\beta} - \hat{\beta})^T(X^TX)(\tilde{\beta} - \hat{\beta})}{ps^2} \tag{5.3}$$

as a distance measure for comparison of least squares on ordinary ridge estimates.

To simplify (5.3), we note that $\tilde{\beta}$ is obtained by adding p fictitious data points (while the estimates considered previously are obtained by deleting points), so a formula for updating regression estimates is required. The required formula can be derived from a result given by Plackett (1965), and

$$D_{R,k} = \frac{\hat{\beta}^TA^T(I + A(X^TX)^{-1}A^T)^{-1}A(X^TX)^{-1}A^T(I + A(X^TX)^{-1}A^T)^{-1}A\hat{\beta}}{ps^2} \tag{5.4}$$

Since the columns of A are $\sqrt{k}$ time the eigenvectors of X^TX, this last equation is simply

$$D_{R,k} = \frac{\hat{\beta}^TA\,\Lambda\,A^T\hat{\beta}}{ps^2} \tag{5.5}$$

Here, Λ is a diagonal matrix with entries $\ell_i/(k + \ell_i)^2$, $i = 1, \ldots, p$ and ℓ_i is the i-th eigenvalue of X^TX. Expression (5.5) makes routine comparison of ridge and least squares estimators for many values of k relatively simple.

6. Computational Considerations

The statistics discussed in this paper are all based on a few elementary building blocks, namely s^2, the residuals, and submatrices and elements of the projection matrix, $V = X(X^TX)^{-1}X^T$; for the one at a time statistics, only the diagonal elements v_{ii} of V are needed. Since most regression programs currently compute residuals, we shall concentrate on the computation of elements of V. Details necessary for actually implementing algorithms may be found in Stewart (1973); also see Seber (1977) for a statistical approach to least squares computations. All matrix operations not specifically given here may be carried out using LINPACK, a set of FORTRAN subroutines from the Argonne National Laboratory (Dongarra, et. al. 1977), to be publicly released in late 1978.

Individual elements of V, the v_{ij}, are easily computed from the Cholesky (square root) factorization of (X^TX). As long as (X^TX) is full rank, we can find a full rank $p \times p$ upper triangular matrix R such that

$$(X^TX) = (R^TR)$$

once R is available, computation of v_{ij} is done using the result

$$\begin{aligned} v_{ij} &= x_i^T(X^TX)^{-1}x_j \\ &= x_i^T(R^TR)^{-1}x_j \\ &= (x_i^TR^{-1})(R^{-T}x_j) \\ &= (R^{-T}x_i)^T(R^{-T}x_j) \ . \end{aligned}$$

To compute v_{ij}, therefore, one needs only to compute $R^{-T}x_i$ and $R^{-T}x_j$; v_{ij} is then the inner product of two p-vectors. Although the triangular matrix R^T is easily inverted, this can be avoided by back substitution (Seber 1977, p.303).

In most applications, it will be convenient to compute and save the v_{ii} whenever the residuals r_i are computed and saved; both can be computed in the same pass through the data. The other one at a time statistics, t_i, F_i, D_i, etc. are then computed using the formulae in earlier sections of this paper.

For the several at a time statistics, the first consideration is obtaining $V_{\underset{\sim}{i}}$, an m×m submatrix of V. Since keeping V, an n×n matrix, explicitly in storage on a computer is often impractical, $V_{\underset{\sim}{i}}$ must be computed for each $\underset{\sim}{i}$ of interest. Obviously, this can be accomplished by repeated use of the method outlined in the last paragraphs: for an m×m submatrix, using 2p registers as scratch to accumulate $R^{-T}x_i$ and $R^{-T}x_j$, m(m-1)/2 back solutions and m(m-1)/2 inner products are required; less computation is possible at the expense of increased storage.

Alternatively, if computations are initially done using the QR factorization, $V_{\underset{\sim}{i}}$ may be easily obtained. Given X of rank p, we can find an n×n orthogonal matrix Q and a p×p upper triangular matrix R such that

$$X = Q \begin{pmatrix} R \\ 0 \end{pmatrix} .$$

If we partition $Q = (Q_1\ Q_2)$, where Q_1 is n×p, then one can show (Seber (1971), p. 304; Stewart (1973), Chapter 7) that

$$V = Q_1 Q_1^T .$$

Thus Q_1 represents a compactification of the n×n matrix as the n×p matrix Q_1, and, if Q_1 can be stored on the machine, a simple algorithm for computing v_{ij} is available. If q_i^T is the i-th row of Q_1, then

$$v_{ij} = q_i^T q_j .$$

Unfortunately, this is a row oriented algorithm and therefore it may be relatively inefficient on some computers.

Once $V_{\underset{\sim}{i}}$ is computed the statistics $F_{(\underset{\sim}{i})}$ and $D_{\underset{\sim}{i}}$ can be computed as well. $F_{(\underset{\sim}{i})}$ is best computed in two steps. First, compute

$$t^2_{\underset{\sim}{i}} = \frac{r_{\underset{\sim}{i}}^T(I-V_{\underset{\sim}{i}})^{-1}r_{\underset{\sim}{i}}}{ms^2}$$

where $a = (I-V_{\underset{\sim}{i}})^{-1}r_{\underset{\sim}{i}}$ is computed by back substitution; since m is usually small, this computation can be carried out be finding the Cholesky factorization of $(I-V_{\underset{\sim}{i}})$. $t^2_{\underset{\sim}{i}}$ is computed as $a^T r_{\underset{\sim}{i}}$ divided by ms^2. $F_{(\underset{\sim}{i})}$ is comuted as

$$F_{(\underset{\sim}{i})} = (n-m-p)t^2_{\underset{\sim}{i}}/(n-p-mt^2_{\underset{\sim}{i}}) .$$

Finally $D_{(\underset{\sim}{i})}$ is computed in a manner analogous to $t^2_{\underset{\sim}{i}}$ by first finding $a = (I-V_{\underset{\sim}{i}})^{-1}r_{\underset{\sim}{i}}$ and then computing the quadratic form $a^T V_{\underset{\sim}{i}} a/ps^2$. If computation of $D_{\underset{\sim}{i}}(\Psi)$ is desired, the necessary modifications are easily worked out.

7. Florida Area Cumulus Experiment (FACE). 1975.

7.1 Description and Initial Considerations

Judging the success of cloud seeding experiments intended to increase rainfall is an important statistical problem. Results from past experiments are mixed. It is generally recognized that, depending on various contributing environmental factors, seeding can produce an increase or decrease in rainfall, or have no effect. Moreover, the critical factors controlling the response are, for the most part, unknown. This fundamental treatment-unit nonadditivity makes judgments about the effects of seeding difficult.

In 1975 the Florida Area Cumulus Experiment (FACE) was conducted to determine the merits of using silver iodide to produce rainfall increases and to isolate some of the factors contributing to the treatment-unit nonadditivity (Woodley et al., 1977). The target consisted of an area of about 3,000 square miles to the north and east of Coral Gables, Florida. In this experiment, 24 days in the summer of 1975 were judged suitable for seeding based on a daily suitability criterion of $S - Ne \geq 1.5$, where S (seedability) is the predicted difference between the maximum height of a cloud if seeded and the same cloud if not seeded, and Ne is a factor which increases with conditions leading to naturally rainy days. (For a more detailed description see Woodley et al., 1977.) Generally, suitable days were those on which the seedability is large and the natural rainfall early in the day is small. On each suitable day, the decision to seed was based on unrestricted randomization; as it happened, 12 days were seeded and 12 were unseeded.

The following variables were measured on each suitable day:

Echo Coverage (C) -- Percent cloud cover in the experimental area, measured using radar in Coral Gables, Florida

Prewetness (P) -- Total rainfall in the target area one hour before seeding (in cubic meters $\times 10^7$)

Echo Motion (E) -- a classification indicating a moving radar echo (1) or a stationary radar echo (2)

Response Variable (Y) -- the amount of rain that fell in the target area for a six-hour period on each suitable day (in cubic meters $\times 10^7$)

The data as presented by Woodley et al. (1977) are reproduced in Table 3. We have also included the variable

Time Trend (T) -- Number of days after the first day of the experiment (June 16, 1975 = 0)

This variable is potentially relevant because there may be a time trend in natural rainfall or modification in the experimental techniques.

In addition to selecting days based on suitability (S - Ne), the investigators attempted to use only days with $C \leq 13$ percent. A disturbed day was defined as $C > 13$. From Table 3, the first two experimental days are disturbed with the second day being highly disturbed (C = 37.9 percent). Because of its very large echo coverage value, it can be anticipated that case 2 will be a high leverage case. We may suspect that the process under study may differ under the conditions of case 2. Therefore, case 2 will be deleted from the primary analysis. The effects of including case 2 will be presented later.

Initially, we shall adopt the model

$$LY = \beta_0 + \beta_1 A + \beta_2 T + \beta_3 (S - Ne) + \beta_4 C + \beta_5 LP + \beta_6 E + \beta_{13}(A \times (S - Ne)) + \beta_{14}(A \times C) + \beta_{15}(A \times LP) + \beta_{16}(A \times E) \quad (7.1)$$

where $LY = \log_{10} Y$ and $LP = \log_{10} P$.

This model contains all linear terms and all cross products between action (A = 1 for seeded days, A = 0 for unseeded days) and the base carriers. The cross product terms are to model the possibility of treatment-unit nonadditivity.

Table 3

Measurements from FACE, 1975

CASE	A	T	S	C	P	E	SA	CA	PA	EA	Y
1	0	0	1.75	13.40	.274	2	0	0	0	0	12.85
2	1	1	2.70	37.90	1.267	1	2.70	37.90	1.267	1	5.52
3	1	3	4.10	3.90	.198	2	4.10	3.90	.198	2	6.29
4	0	4	2.35	5.30	.526	1	0	0	0	0	6.11
5	1	6	4.25	7.10	.250	1	4.25	7.10	.250	1	2.45
6	0	9	1.60	6.90	.018	2	0	0	0	0	3.61
7	0	18	1.30	4.60	.307	1	0	0	0	0	.47
8	0	25	3.35	4.90	.194	1	0	0	0	0	4.56
9	0	27	2.85	12.10	.751	1	0	0	0	0	6.35
10	1	28	2.20	5.20	.084	1	2.20	5.20	.084	1	5.06
11	1	29	4.40	4.10	.236	1	4.40	4.10	.236	1	2.76
12	1	32	3.10	2.80	.214	1	3.10	2.80	.214	1	4.05
13	0	33	3.95	6.80	.796	1	0	0	0	0	5.74
14	1	35	2.90	3.00	.124	1	2.90	3.00	.124	1	4.84
15	1	38	2.05	7.00	.144	1	2.05	7.00	.144	1	11.86
16	0	39	4.00	11.30	.398	1	0	0	0	0	4.45
17	0	53	3.35	4.20	.237	2	0	0	0	0	3.66
18	1	55	3.70	3.30	.960	1	3.70	3.30	.960	1	4.22
19	0	56	3.80	2.20	.230	1	0	0	0	0	1.16
20	1	59	3.40	6.50	.142	2	3.40	6.50	.142	2	5.45
21	1	65	3.15	3.10	.073	1	3.15	3.10	.073	1	2.02
22	0	68	3.15	2.60	.136	1	0	0	0	0	.82
23	1	82	4.01	8.30	.123	1	4.01	8.30	.123	1	1.09
24	0	83	4.65	7.40	.168	1	0	0	0	0	.28

A = Action (0 = not seeded, 1 = seeded)

T = Time in days (June 16, 1975 = 0)

S = S-Ne = Seeding Suitability Criterion

C = Echo Coverage in Percent

P = Prewetness (in cubic meters $\times 10^7$)

E = Echo Motion (1 = moving radar echo, 2 = stationary radar echo)

SA = $(S-N_e) \times A$

CA = $C \times A$

PA = $P \times A$

EA = $E \times A$

Y = Rainfall (in cubic meters $\times 10^7$)

Because of the limited degrees of freedom available, higher order terms in any of the base carriers have not been included. The response variable and prewetness were transfored to logarithms because these are volume measures whereas others (eg. S-Ne) are linear and because we expect non-constant variances in the original scale. (Residual plots in the untransformed scale confirm the need for a transformation.)

The main goal of our analysis is to describe the difference, ΔLY, between predicted rainfall for seeded days and unseeded days,

$$\Delta LY = LY(A = 1) - LY(A = 0) = \beta_1 + \beta_{13}(S - Ne) + \beta_{14}C + \beta_{15}LP + \beta_{16}E. \quad (7.2)$$

Thus, the additive effect β_1 and the four possible interaction terms are of primary interest. The prediction of rainfall by itself is of secondary interest. Table 4 gives the least square estimates and their estimated standard errors for the coefficients in (7.1) as well as the estimate for a few selected subset models to be discussed later. From the first column of Table 4, β_1, β_2, β_6 and β_{13} are apparently distinguishable from zero, while others appear unnecessary. However, before the model is refined, the impact of each case on the estimated coefficients needs to be assessed. Because of the special interest in the coefficients of (7.2), we consider both the overall distance measure D_i and the distance meansure $D_i(\psi)$ corresponding to the subset $\psi^T = (\beta_1, \beta_{13}, \beta_{14}, \beta_{15}, \beta_{16})$.

7.2 Case Analysis: Full Model

Table 5 gives r_i, t_i, v_{ii}, $F_i^{1/2}$ (see 2.12), D_i and $D_i(\psi)$ for the full model without case 2. The largest two values of each statistic except v_{ii} listed in Table 5 correspond to cases 7 and 24, both unseeded days with unusually low rain fall recorded. The values of $F_i^{1/2}$ for i = 7,24 and the

Table 4

Estimated Coefficients, Standard Errors and Root Mean Squared Error (RMSE) for Selected Data Sets and Models

Coefficient	Cases Deleted (2)	(2,7,24)	(2,7,24)	(2,7,24)	(7,24)
$\hat{\beta}_0$	-0.291[a] (0.498)[b]	0.417 (0.400)	0.492 (0.129)	0.436 (0.145)	0.400 (0.142)
$\hat{\beta}_1$(A)	2.244 (0.819)	1.426 (0.510)	1.294 (0.190)	1.381 (0.216)	1.458 (0.206)
$\hat{\beta}_2$(T)	-0.009 (0.003)	-0.006 (0.002)	-0.007 (0.001)	-0.006 (0.001)	-0.006 (0.001)
$\hat{\beta}_3$(S-N_e)	0.136 (0.114)	0.006 (0.085)	*	*	*
$\hat{\beta}_4$(C)	0.025 (0.028)	0.030 (0.015)	0.022 (0.010)	0.028 (0.012)	0.030 (0.012)
$\hat{\beta}_5$(LP)	0.436 (0.266)	0.341 (0.146)	0.399 (0.083)	0.379 (0.087)	0.357 (0.085)
$\hat{\beta}_6$(E)	0.573 (0.261)	0.265 (0.135)	0.301 (0.074)	0.295 (0.074)	0.293 (0.075)
$\hat{\beta}_{13}$(A×S-N_e)	-0.465 (0.178)	-0.333 (0.107)	-0.326 (0.052)	-0.319 (0.053)	-0.309 (0.053)
$\hat{\beta}_{14}$(A×C)	-0.011 (0.057)	-0.023 (0.028)	*	-0.021 (0.024)	-0.045 (0.012)
$\hat{\beta}_{15}$(A×LP)	-0.049 (0.443)	0.073 (0.224)	*	*	*
$\hat{\beta}_{16}$(A×E)	-0.291 (0.354)	0.050 (0.178)	*	*	*
RMSE ($\hat{\sigma}$)	0.291	0.139	0.122	0.123	0.124

[a] Estimated coefficient

[b] Estimated standard error

* Term omitted from computations

studentized residual plot given in Figure 2 suggest that these cases do not conform to the assumed model. Using a Bonferroni inequality, the p-value corresponding to the outlier test for case 7 is 0.055. The most likely candidate for a pair of outliers is (7,24) and the associated F-statistic can be computed to be 21.21 on 2 and 10 degrees of freedom. The Bonferroni p-value is 0.064. Although (7,24) is evidently an outlier pair, it has little influence on the least squares estimate of β or ψ, $D_{(7,24)} = 0.455$. and, removal of (7,24) will move $\hat{\beta}$ only to the edge of a 10% confidence ellipsoid.

As an alternative to deleting cases 7 and 24, (that is, modeling them with indicator vectors) we could consider attempting to expand the model to include additional terms in the base carriers. For example, if a variable $(S\text{-}Ne)^2$ is added to the model, then the residuals for both cases 7 and 24 becomes relatively small. However, the influence measure for these two cases on the parameter estimate for $\psi^* = \{(S\text{-}Ne)^2\}$ is $D_{7,24}(\psi^*) = 19.71$, which suggests that including this variable has little effect other than providing an alternative model for cases 7 and 24. While in this problem we prefer to delete these cases, in other problems adding a different variable may be preferable.

The most influential pair of observations is (3,20) , $D_{(3,20)} = \infty$; that is, when these observations are removed the model becomes rank deficient and a unique least squares estimate of β does not exist. The deficiency arises because A and E × A are idential after cases 3 and 20 are removed. This situation could also have been detected by noting that in the raw data cases 3 and 20 are the only ones with a "2" in the E × A column, although in large data sets such examinations become impractical. Alternatively, an inspection of the correlations between the residuals would have revealed the problem since the residual correlation for (3,20) is -1.0. Although cases 3 and 20

TABLE 5

Univariate Case Statistics for the Full

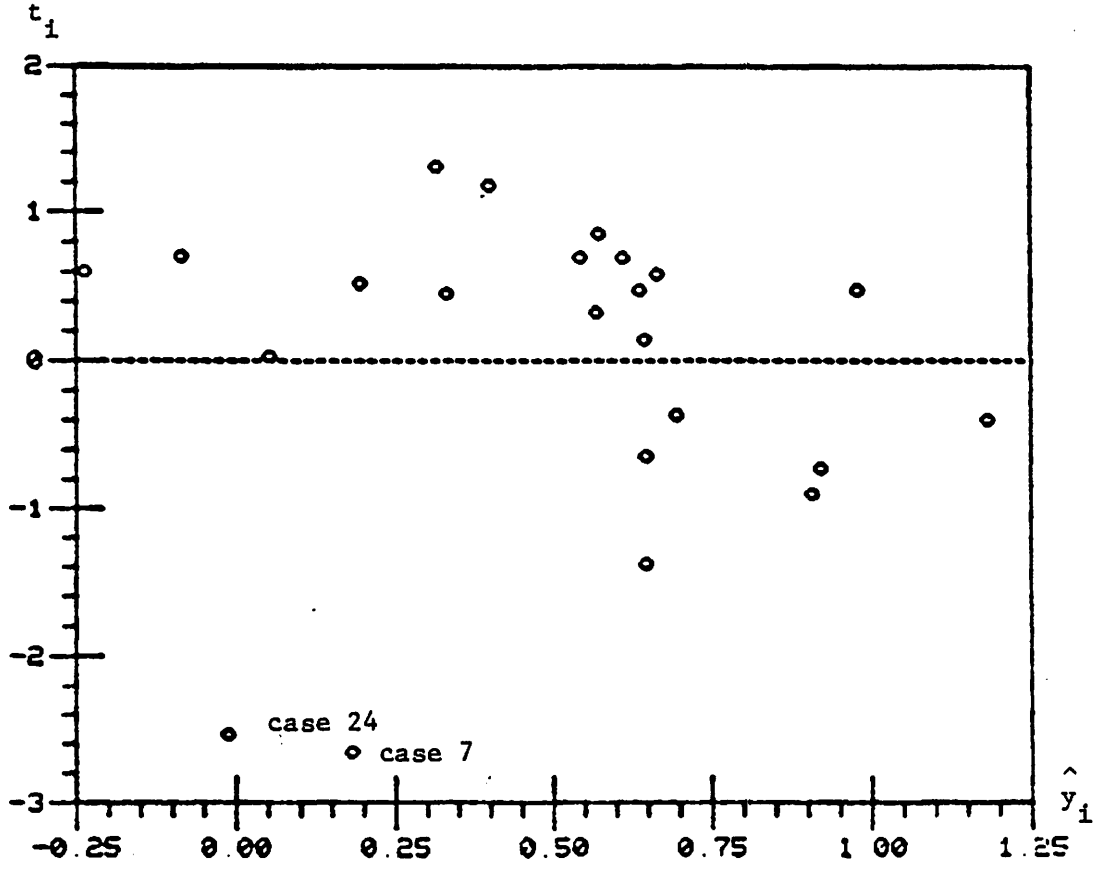
Model with Case 2 Deleted

CASE(i)	r_i	t_i	v_{ii}	$F_i^{1/2}$	D_i	$D_i(\psi)^*$
1	-.072	-.383	.580	.369	.018	.008
3	-.124	-.714	.646	.699	.085	.052
4	.211	.872	.307	.863	.031	.018
5	-.258	-1.366	.578	1.423	.232	.206
6	.158	1.190	.793	1.213	.492	.357
7	-.510	-2.643	.560	3.916	.810	.737
8	.343	1.322	.208	1.370	.042	.050
9	.138	.604	.386	.587	.021	.009
10	-.204	-.889	.379	.880	.044	.052
11	.110	.468	.354	.452	.011	.010
12	-.090	-.352	.234	.338	.003	.004
13	.121	.492	.286	.475	.009	.008
14	.039	.154	.256	.148	.001	.001
15	.094	.497	.583	.481	.031	.050
16	.079	.339	.365	.326	.006	.004
17	-.086	-.622	.777	.606	.123	.144
18	.081	.713	.848	.698	.257	.363
19	.010	.042	.261	.040	.000	.000
20	.124	.714	.646	.699	.085	.099
21	.109	.543	.529	.527	.030	.043
22	.149	.614	.301	.597	.015	.014
23	.121	.720	.669	.704	.095	.116
24	-.541	-2.519	.455	3.512	.482	.380

* $\psi^T = (\beta_1, \beta_{13}, \beta_{14}, \beta_{15}, \beta_{16})$

Figure 2

Studentized Residuals, t_i, vs. Predicted Values, $\hat{y}_i$, from the Full Model with Case 2 Deleted



are jointly influential, they are individually uninfluential and there is no apparent reason to doubt their authenticity. No other pair has a serious influence on $\hat{\beta}$ or $\hat{\psi}$.

At this point, we choose to delete pair of suspected outliers, (7,24). The second column of Table 4 gives the estimated coefficients for the full model without (2,7,24). Note that three of the estimated coefficients are negative. $\hat{\beta}_1$ is approximately half of its previous value. Table 6 gives the univariate case statistics for the full model without (2,7,24) and the corresponding studentized residual plot is given in Figure 3. The residual plot is well-behaved and the univariate case statistics reveal no problems or anomalies. Inspection of the pairwise case statistics reveals no joint outliers and, of course, (3,20) is still the most influential pair. However, in addition, there is a second pair which is highly influential, $D_{(4,16)} = 7.1$ and $D_{(4,16)}(\psi) = 9.0$. Individually these cases are uninfluential. The high joint influence appears to be the result of a large residual correlation, +0.89. Since the residual correlation is positive, cases 4 and 16 probably lie on opposite sides of the center of the data, nearly on the same ray (Cook, 1979). While we have identified the fact that removing these cases may lead to substantially different conclusions, we have no real need or justification for doubting the usefulness of these two cases. Our strategy is to leave them in for further analyses, and at later steps continue to monitor their influence.

7.3 Final Model (cases 2,7 and 24 deleted)

The final model was selected by calculating all possible subset regressions (Furnival and Wilson (1974)) and selecting the best 5 using Mallow's C_p criterion as an approximation to a mean square error of prediction criterion (Bingham, Cook and Weisberg (1978)). Two final models were chosen. The third

TABLE 6

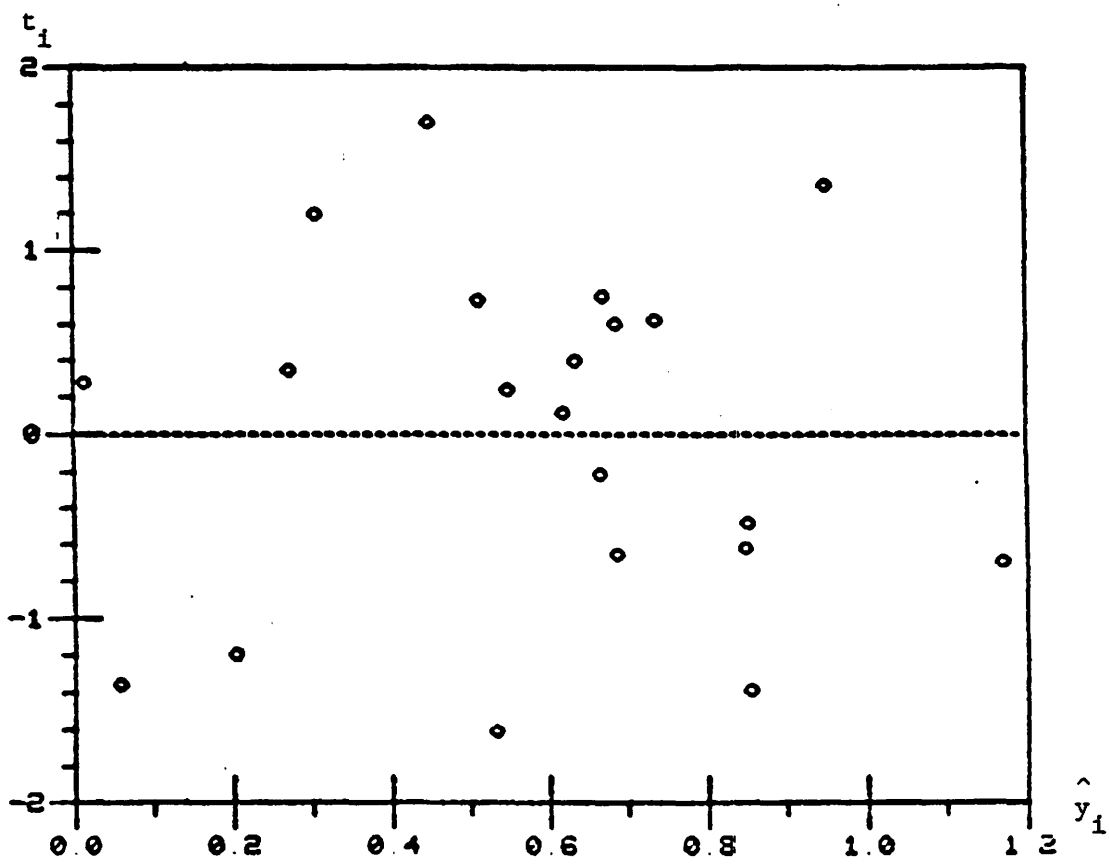
Univariate Case Statistics for the Full Model with Cases 2,7 and 24 Deleted.

CASE(i)	r_i	t_i	v_{ii}	$F_i^{1/2}$	D_i	$D_i(\psi)^*$
1	-.060	-.688	.606	.668	.066	.027
3	-.050	-.610	.653	.590	.064	.037
4	.051	.633	.670	.613	.074	.080
5	-.142	-1.598	.594	1.756	.340	.289
6	.045	.741	.809	.723	.212	.159
8	.209	1.714	.234	1.935	.082	.105
9	-.049	-.470	.436	.451	.016	.008
10	-.150	-1.374	.382	1.447	.106	.112
11	.135	1.208	.355	1.240	.073	.062
12	-.078	-.642	.234	.622	.011	.012
13	.089	.764	.303	.747	.023	.023
14	.049	.410	.256	.392	.005	.006
15	.123	1.370	.584	1.442	.239	.347
16	-.017	-.207	.648	.196	.007	.008
17	.015	.244	.805	.232	.022	.027
18	.006	.119	.855	.113	.008	.010
19	-.138	-1.184	.297	1.212	.054	.064
20	.050	.610	.653	.590	.064	.070
21	.034	.358	.536	.342	.013	.018
22	-.144	-1.345	.409	1.410	.114	.135
23	.023	.287	.681	.274	.016	.012

* $\psi^T = (\beta_1, \beta_{13}, \beta_{14}, \beta_{15}, \beta_{16})$

Figure 3

Studentized Residuals, t_i, vs. Predicted Values, $\hat{y}_i$, from the Full Model with Cases 2,7 and 24 Deleted



column of Table 4 gives the estimated coefficients for the model we consider "best" and the fourth column gives those for a possible second choice. The two models differ by the presence of the A×C(β_{14}) term. The estimates of the coefficients in common to the two models are quite close. Also, note that $|\hat{\beta}_{14}|$ is less its standard error and, thus, contributes little and might be judged unnecessary.

The univariate case statistics for the "best" model are given in Table 7. The univariate case statistics and residual plot all appear well-behaved. The bivariate cases statistics were inspected also and again no problems were noted. In particular, (3,20) and (4,16) are no longer influential: The E×A term has been deleted and the residual correlation for (4,16) is now -0.06.

As a check on the influence of (3,20) and (4,16), the best 5 models using the C_p criterion were computed without various combinations of these points; our final model was always among the best 5. Evidently, these points have little influence on the terms in the final solution.

In the final model, the estimated predicted difference $\hat{\Delta LY}$ (see 7.2) contains the seeding effect and only one of the four possible interaction terms,

$$\hat{\Delta LY} = 1.29 - .33(S\text{-}Ne) \quad . \tag{7.3}$$

Note that the coefficient of the seeding suitability criterion, S-Ne, is negative and, thus, the predicted difference in rainfall <u>decreases</u> as S-Ne increases. According to this result, <u>seeding produces a decrease in rainfall</u> when S-Ne > 3.91. A plot of (7.3) along with the 95% simultaneous prediction bands computed using the Scheffe method is given in Figure 4.

In short, contrary to the experimenters' prior opinion, there is evidence to suggest that optimal seeding occurs if the seeding suitability criterion is low!

TABLE 7

Univariate Case Statistics for the Final Model with Cases 2,7 and 24 Deleted

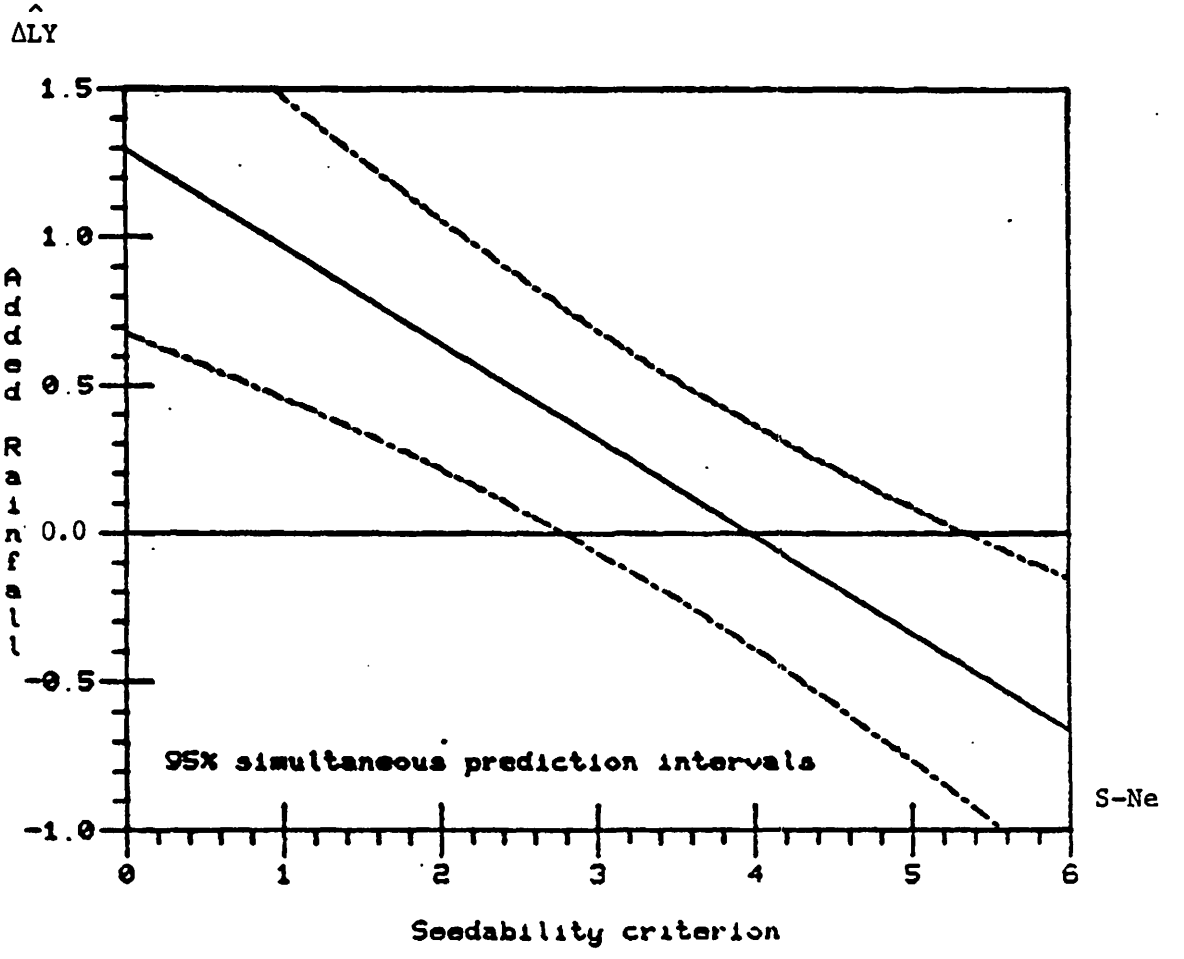
CASE(i)	r_i	t_i	v_{ii}	$F_i^{1/2}$	D_i	$D_i(\psi)^*$
1	-.054	-.592	.438	.577	.039	.002
3	-.037	-.407	.446	.395	.019	.013
4	.015	.141	.278	.136	.001	.001
5	-.187	-1.934	.372	2.177	.316	.596
6	.068	.915	.628	.910	.202	.112
8	.208	1.881	.176	2.097	.108	.178
9	-.026	-.263	.318	.254	.005	.000
10	-.165	-1.673	.351	1.802	.216	.405
11	.142	1.393	.304	1.446	.121	.296
12	-.051	-.451	.158	.438	.005	.005
13	.075	.689	.197	.675	.017	.011
14	.072	.642	.164	.628	.012	.010
15	.090	.957	.405	.954	.089	.233
16	.026	.251	.263	.242	.003	.001
17	-.022	-.233	.422	.225	.006	.003
18	.045	.453	.348	.440	.016	.008
19	-.151	-1.430	.253	1.492	.099	.159
20	.045	.473	.401	.459	.021	.014
21	.062	.577	.214	.563	.013	.002
22	-.140	-1.377	.310	1.427	.122	.193
23	-.016	-.199	.556	.192	.007	.008

* $\psi^T = (\beta_1, \beta_{13})$

Figure 4

Predicted Difference $\hat{\Delta LY}$ vs. S-Ne

from the Final Model



7.4 Case 2

Recall that case 2 was deleted at the outset on the grounds that it would probably be influential and that it may not conform to the process under study. The F-statistic for testing the fit of case 2 to the final model of section 7.3 has the value 16.00 and 1 and 16 degrees of freedom and the associated p-value is 0.001. However, it is possible that case 2 can be explained by one of the deleted terms. To check on this possibility, the previous analysis was repeated with case 2 included.

All qualitative conclusions reached in the analysis without case 2 remain valid with case 2. Also, as expected, case 2 is highly influential. For example, in the full model after deleting the outlier pair (7,24) the distance measure for case 2 is $D_2 = 3.25$.

The primary difference between the two analyses is in the final models. The last column of Table 4 gives the estimated coefficients for the model judged best using C_p when case 2 is present. A comparison of the last three columns of Table 4 suggests that case 2 is influential for only the A×C term. This is confirmed by the distance measures for the subsets (β_{14}) and $\psi_1^T = (\beta_0, \beta_1, \beta_2, \beta_4, \beta_6, \beta_{13})$ from the final model, $D_2(\beta_{14}) = 4.15$ and $D_2(\psi_1) = 0.57$. Evidently, the A×C term is needed to model case 2 only.

The predicted difference from the final model with case 2 is

$$\hat{\Delta LY} = 1.46 - 0.31(S\text{-}Ne) - 0.045C \qquad (7.4)$$

This suggests, in addition to previous conclusions, that the effect of seeding decreases with increasing cloud cover. Admittedly, this conclusion seems odd.

Finally, the full model with case 2 was fit using Huber's proposal 2 (1973) robust estimator with a variety of truncation points. The scale was chosen as median $[|r_i|/0.6745]$ and Bickel's proposal 2 (1975) was used as

the stepping method with Andrew's median estimate (1974) as a starting value. The results are generally consistent with previous conclusions. For example, the estimated predicted difference after 15 iterations with truncation point 1.0 is

$$\hat{\Delta LY} = 1.53 - 0.31(S\text{-}Ne) - 0.045C - 0.061LP - 0.053E \quad .$$

and the scale estimate is 0.127.

References

Andrews, D. F. (1974), "A robust method for multiple linear regression," Technometrics 16, 523-32.

Andrews, D. F. et al. (1972), Robust Estimate of Location. Princeton: Princeton Univ. Press.

Andrews, E. F. and Pregebon, D. (1978), "Finding outliers that matter," Journal of the Royal Statistical Society Ser. B 40, 85-93.

Bickel, P. (1975), "One-step Huber estimates in the linear model," Journal of the American Statistical Association 70, 428.

Bingham, C. (1977), "Some identities useful in the analysis of residuals from linear regression." University of Minnesota, School of Statistics, Technical Report No. 300.

Bingham, C., Cook, R. D. and Weisberg, S. (1978), "A mean square error criterion for subset selection," Submitted for publication.

Cook, R. D. (1977), "Detection of influential observations in linear regression," Technometrics 19, 15-18.

Cook, R. D. (1979), "Influential observations in linear regression," Journal of the American Statistical Association (in press).

Dongarra, J. J. et al. (1977), Preliminary LINPACK User's Guide. Argonne National Laboratory, LINPACK Working Note #9.

Furnival, G. and Wilson, R. (1974), "Regression by leaps and bounds," Technometrics 16, 499-511.

Gentleman, J. F. and Wilk, M. B. (1975), "Detecting outliers II: Supplementing the direct analysis of residuals," Biometrics 31, 387-410.

Hampel, F. R. (1974), "The influence curve and its role in robust estimation," Journal of the American Statistical Association 69, 383.

Hoaglin, D. C. and Welsch, Roy (1978), "The hat matrix in regression and ANOVA," American Statistician 32, 17-22.

Huber, P. J. (1972), "Robust statistics: A review," Annals of Mathematical Statistics 43, 1041-78.

Huber, P. J. (1973), "Robust regression: Asymptotics, conjectures and Monte Carlo," Annals of Statistics 1, 799-821.

Seber, G.A.F. (1977), Linear Regression Analysis. New York: Wiley.

Stewart, G. W. (1973), Introduction to Matrix Computation. New York: Academic Press.

Welsch, R. E. and Kuh, E. (1977), "Linear Regression Diagnostics," Working Paper No. 173, National Bureau of Economic Research, Cambridge, Mass.

Woodley, W. L. et al. (1977), "Rainfall results, 1970-75: Florida area cumulus experiment," Science 195 (25 Feb 1977), 735-742.